Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.04:004.65

Ибрахим Максим Мохамад

Исследование алгоритмов для обработки больших данных

АВТОРЕФЕРАТ
на соискание академической степени магистра
по специальности 1-40 80 02 Системный анализ, управление и обработка информации (Системный анализ и управление в технических системах)

_____

Научный руководитель

Петровский Иосиф Иосифович

Кандидат технических наук, доцент

_____

Минск 2021 г

ВВЕДЕНИЕ

Обширный спрос на услуги обостряет борьбу за клиентов, содействует поиску конкурентных преимуществ, влечет за собой более оперативную реакцию на запросы потребителей. Именно, вследствие этого предприятия все больше стремятся инвестировать свои финансовые средства в инновационные информационные технологии, особенно в технологии работы с БД.

Технологии БД содействуют успешному решению большого количества сложных и актуальных задач: от анализа отзывов о бренде и до ориентации на клиентов; от приверженности требованиям регулятора и до необходимости осуществления усовершенствования базовых систем. Внедрения решений в сферы больших данных сегодня приобретают большую популярность среди различных предприятий, и обеспечивают им значительную выгоду в практически всех направлениях их деятельности. Представить крупное предприятие без современных технологий по работе с базами данных, как минимум в трех из направлений уже практически невозможно:

- Соблюдение требований законодательства и администрирование рисков. Технологии БД могут использоваться для мониторинга поведения клиентов с целью раскрытия подозрительной активности, увеличения точности данных, роста производительности, уменьшения количества ошибок, оперативного реагирования на претензии клиентов, предотвращения мошенничества.
- Увеличение качества обслуживания клиентов. Технологии БД могут применяться, чтобы улучшить взаимодействие предприятий с их клиентами, спрогнозировать отзывы на маркетинговые кампании и для оценки их результатов, а также для персонального подхода к каждому клиенту и поощрения добросовестных клиентов.
- Увеличение операционной эффективности. Самое последнее поколение аналитики может применяться в целях обработки неструктурированных потоков данных, позволяя обеспечить визуализацию проходящих бизнес-процессов, событий и операций, для выполнения мониторинга актуализации информации на административных панелях, своевременной рассылки информации.

Бесспорно, технологии БД положительно сказываются на работе любого предприятия, и эффект от их использования будет повышаться по мере того, как решения в области баз данных будут становиться все более совершенными. Предполагается, что предприятия в ближайшей перспективе будут наращивать объемы накапливаемых данных, использовать новые источники и применять новые сценарии для обрабатывания и последующего анализа данных. У предприятий будет появляться интерес к поиску новых моделей взаимодействия с остальными

сферами (к примеру, телекоммуникации и торговли), что позволит им делать своевременные целевые предложения своим клиентам.

Объектом в работе выступает технология Больших данных. В качестве предмета исследования рассматривается процесс проверки качества в данных технологиях.

Цель исследования состоит в исследовании и разработке системы проверки качества данных для BigData.

Исходя из поставленной цели необходимо выполнить некоторые задачи:

- проведение теоретических исследований в сфере анализа данных;
- рассмотрение информационных систем в целях проверки качества данных для BigData и осуществление выбора наиболее соответствующей анализируемому объекту;
- исследование и выбор методов оценки качества данных для BigData;
- осуществить разработку модели внедрения и модели оптимизации информационных систем качества данных для BigData;
- описать и проанализировать архитектуру внедряемой системы;
- дать описание принципа работы системы и выполнить оценку качества данных для BigData;
- осуществить апробацию работы.

Научное новшество работы состоит в разработке моделей проверки качества данных для BigData.

Практическое значение состоит во внедрении системы проверки качества данных для BigData.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Тема выпускной квалификационной работы: «Исследование алгоритмов для обработки больших данных».

Объектом в работе выступает технология Больших данных. В качестве предмета исследования рассматривается процесс проверки качества в данных технологиях.

Цель исследования состоит в исследовании и разработке системы проверки качества данных для BigData.

В работе рассматривается внедряемая система, ее преимущества, функции, а также осуществлена разработка ее архитектуры. Внедряемый программный продукт прошел апробацию на реальных статистических данных, предоставленных немецкой компанией для проверки качества данных. Построены две модели проверки качества данных: с применением коэффициента WOE для трансформации категориальных переменных в числовые, и с применением порядковой нумерации.

Выполнен сравнительный анализ способа логистической регрессии, который реализован в SAP Analytical Studio, со способами деревьев решений и картой проверки качества данных, построенными в системе SAP Analytics. Наилучший результат по критерию общей точности модели продемонстрировал способ деревьев решений, а индекс Gini был наилучшим в карте проверки качества данных. Логистическая регрессия в SAP Analytical Studio по обоим критериям продемонстрировала средние результаты, что свидетельствует о ее хорошей прогнозирующей способности.

Практическое значение состоит во внедрении системы проверки качества данных для BigData.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались на следующих научных конференциях:

– 21 международная научная конференция “Сахаровские чтения 2021 года” (институт имени А.Д. Сахарова, 2021)

СОДЕРЖАНИЕ РАБОТЫ

Суть работы заключается в исследовании и разработке системы проверки качества данных для BigData.

Исходя из поставленной цели необходимо выполнить некоторые задачи:

- проведение теоретических исследований в сфере анализа данных;
- рассмотрение информационных систем в целях проверки качества данных для BigData и осуществление выбора наиболее соответствующей анализируемому объекту;
- исследование и выбор методов оценки качества данных для BigData;
- осуществить разработку модели внедрения и модели оптимизации информационных систем качества данных для BigData;
- описать и проанализировать архитектуру внедряемой системы;
- дать описание принципа работы системы и выполнить оценку качества данных для BigData;
- осуществить апробацию работы.

Научное новшество работы состоит в разработке моделей проверки качества данных для BigData.

Практическое значение состоит во внедрении системы проверки качества данных для BigData.